Comparison of Machine Learning Algorithms for Mapping the Phytophysiognomies of the Brazilian Cerrado

Luciano T. de Oliveira[1], Thomaz C. de A. Oliveira[2], Luis M. T. de Carvalho[3], Wilian Soares Lacerda[4], Samuel R. de S. Campos[5], Adriana Z. Martinhago[6]

[1]Departamento de Ciências Florestais – Universidade Federal de Lavras (UFLA)
Caixa Postal 3.037 – 37.200-000 – Lavras – MG – Brazil

oliveiralt@yahoo.com.br, thomaz@vialavras.com.br, passarinho@ufla.br, lacerda@ufla.br, samuelcampos@ufla.br, dricazn@gmail.com

Abstract. *This present work describes the classification of the Phytophysiognomies present in the Brazilian Cerrado biome through the means Artificial Intelligence; data from remote sensing images and other sources served as input for these algorithms to generate the vegetation maps. The data acquired was of many types so that it fully described the various Phytophysiognomies present in biome and served as training data for the machine learning algorithms. Various statistical and neuro-computation based algorithms were used for pattern recognition in the data so that we could build a good generalization model for the biome. A vegetation map was successfully generated with each algorithm. Finally a comparison among these algorithms was made so that we could find the best algorithm that fitted the problem of mapping this biome.*

keywords: Image classification, decision trees, maximum likelihood, neural network, Cerrado Phytophysiognomies

1. Introduction

The cerrado biome of tropical South America covers about 2 million km^{2,} an area approximately the same as that of Western Europe, representing ca. 22% of the land surface of Brazil. The biome was named after the vernacular term of its predominant vegetation type a fairly dense woody savanna of shrubs and smalls trees. The term Cerrado (Portuguese for "half-closed," "closed" or "dense") was probably applied to this vegetation originally because of the difficulty of traversing it on horseback. (Oliveira-Filho et al 2002) The constant threat to the Brazilian Cerrado has lead to the necessity of strategies and measures to promote the monitoring and mapping of this biome. The Cerrado has a large biodiversity but it's fragmentation throughout the years has lead to the losses of exemplars from this biome. This process can be noticed in the red books of fauna (Machado et al., 1998) and flora (Mendonça & Lins, 2000) of the Minas Gerais.

The absence of a precise mapping occurs not only of this biome but to the others present in the state of Minas Gerais too. This leads to difficulties in the environmental

management due to database deficiencies reflecting in other aresas management of the state.

This can be noticed in the northern part of the Minas Gerais state, where the biome of Cerrado occurs very intensively. In this region we can observe areas with big social problems that intensify by the lack of social and forest management that lead to a clandestine exploration of vegetable coal which is intensified by the product's high market value.

This work's main objective is to develop an efficient methodology to generate the mapping of the various phytophysiognomies present in the Savanna biome region and to promote a comparison among the classification algorithms used to generate theses vegetation maps. Precise vegetation mapping can help the monitoring and the environmental administration of such areas. The main objective of this work is to propose a methodology based on machine learning algorithms that can help achieve this objective of precise mapping of the phytophysiognomies present in the Cerrado biome.

2. Methods

2.1. Field sampling of the phytophysiognomies.

Initially the classification proposed by Ribeiro & Walter (1998) was used to promote the characterization and for the choice of division level of the phytophysiognomies in the Cerrado biome.

Throughout qualitative analyses of images EMT+, during a low humidity period and a high humidity period, some areas were identified as representative areas of forest fragments and also components of the agricultural landscape (Louzada, 2000).

Some field expeditions by air and land occurred for identification and analyses of the distribution of the remaining phytophysiognomies of the region; these were latterly used as samples of earth observation trueness.

The phytophysiognomical levels were adapted from the original classification from Ribeiro & Walter (1998) because of the necessity to adequate the characteristics of the analyzing sensor that is not capable of resolving small forest fragments that are representative of one phytophysiognomy. Valey-side marshy grasslands fragments are very small and tend to be found mixed among riverine forests. Open grassland and grassland with scattered shrubs fragments are very associated to each other. By the characteristics of the fragments just described, some of the fragments individually would not be captured because of their small size and the object's size must be at least three times the size of the sensor's resolution.

Analyzing an individual date of remote sensing data to extract meaningful vegetation biophysical information is often of value. However timing is a very important when attempting to identify different vegetation types or to extract useful vegetation biophysical information (e.g. biomassa, chlorophyll characteristics) from remotely sensed data (Jensen 2000). The group of samples was initially established from the EMT image from a dry spell period/high humidity period. From the entire group of samples, 30% were separated

randomly from each class of phytophysiognomies. These samples were used as training data for the algorithms, accuracy data of the classification and latter as test data of the accuracy (**Table 1**). This last one being only used on the decision tree algorithms. By this way, was a compromise between stratified sampling and randomly chosen sampling was set.

TABLE 1 - Total number of sampled pixels

	Forests	**Savannic**	**grassland**	**Eucalipte plantation**	**pasture**	**cropland**	**Bair soil**	**Water**	**Shades**
Samples	8237	67251	53956	18710	42396	1814	38487	5635	2104

2.2. Image Processing

With the goal of reducing the noise resulting form image fusion, and due to the atmospheric influence in the panchromatic image, the Lee filter with a 3x3 window was used so that it would reduce the texture resulting from the noise but not the image detail.

Vegetation indices are dimensionless, radiometric measures that function as indicators of relative abundance and activity of green vegetation. A vegetation index should: maximize sensitivity to plant biophysical parameters, normalize external effects such as Sun angle, normalize internal effects such as canopy background variations. There are more than 20 vegetation indices in use (Jensen 2000). A NDVI (Normalized Difference Vegetation Index) was calculated in all of the images, making a relationship in the band that represents the red and infrared wavelengths bands, those bands correspond to the bands 3 and 4 if the EMT+ images.

The Tasseled Cap transformation is a global vegetation index. Theoretically, it may be used anywhere in the world to disaggregate the amount of soil brightness, vegetation, and moisture content in individual pixels in a Landsat MSS or Thematic Mapper image (Jensen 2000). The coefficients necessary for the Tasseled Cap (**Table 2**) were only applied on the ETM+ images. (Crist & Cicone 1984) calculated these coefficients so that they can be applied on digital images.

TABLE 2 - Coefficients of Tasseled Cap applied on Landsat images

Indexes	**ETM+ 1**	**ETM+ 2**	**ETM+ 3**	**ETM+ 4**	**ETM+ 5**	**ETM+ 7**
Brightness	0,3037	0,2793	0,4743	0,5585	0,5082	0,1863
Greenness	-0,2848	-0,2435	-0,5436	0,7243	0,0840	-0,1800
Wetness	0,1509	0,1973	0,3279	0,3406	-0,7112	-0,4572

The mixture fractions were obtained taking into account the simplex theory (Correia, 1983, Aguiar, 1991; Mather, 1999; Tso & Matter 2001, Schowengerdt, 1997), thus obtaining the pure pixels from the extremes of the distribution from the sampling space domain (Red X infrared).

According to (INPE 2002), there is no necessity to convert the digital pixels into values of reflectance when the values were obtained from the image itself. Thus it was decided to leave the values in digital numbers.

The model was applied with some restrictions; the fractions of shades, vegetation and soil could not overcome 100% of total mixture found in a pixel. One image of 25 classes was generated by the ISODATA unsupervised classification method, applying 10 iterations with a minimum value of 10 pixels per class and gathering a number of six isolated pairs in the class. This image was created with the intention of simplifying the information of the image, helping in this case the performance of classifying algorithms and the distinction of the each pixel in the classification. This is helpful, once the classifiers work on every pixel individually.

With altitude curves in the forms of vectors from IBGE institute, it was generated one image of the classes of altitude. In this image the altitude was rearranged into a 0 to 255 range, the lowest altitude being 0 and the highest altitudes of the region 255.

For the river buffer it was necessary the extraction of the whole of the hydrography through a visual analyses of high resolution images. The buffer ranges from 0 up 255, zero being the value where river stands and it gradually increases up to 255 as the distance increases from the river. Values of 255 correspond to locations distant from a river pixel.

The images of classes of altitude and hydrography are very important to this work since they establish import features and relationships that characterize the vegetation of the Cerrado biome.

2.3. Image classification

As commented earlier, the main objective of this work is to compare image classification algorithms among themselves. For Moreira (2003) an automatic image identification and classification can be sought as the analyses and the manipulation of images through computational techniques, with the goal of extracting information regarding an object of the real world. For this research, maps of the phytophysiognomies of the Cerrado biome were generated with the following algorithms: Decision trees, Maximum likelihood, Kohonen's self Organizing maps with supervised learning, Multi layer perceptrons and Fuzzy ART maps Neural networks.

2.3.1. Maximum Likelihood

The Maximum likelihood is a statistical algorithm that necessitates some previous sampling before its operation (learning stage of the classifier), where it can be established a previous indication of the number and a specific pattern of a certain class. (Lillesland & Kiefer, 2000).

This classifier is based in the Bayesian theory of probability; it uses an array of patterns and a covariance matrix from a Gaussian distribution sample set. (Lillesland & Kiefer, 2000, Gonzáles e Woods, 2000 ; Tso & Mather, 2001). The classification is therefore

defined by the smallest number of standard deviation from sample set. Thus each pixel is classified according to an average array and covariance matrix. The maximum likelihood distinguishes from the other classifiers by having good overall performance for classifying Earth's surface. (Carvalho, 2001 ; Marcelino et al., 2003 ; Oliveira et al., 2002). 30% from the total number of samples of earth trueness were used as training data set of the algorithms; these values are present in **Table 1**. In the maximum likelihood it was considered that all the pixels had the same probability of belonging to each one of the present classes.

2.3.2. Decision Tree

The decision tree is a non parametrical classifier that is based in the inductive learning of a human being, where one can learn to separate the classes throughout training data (Quilan, 1986). From the training data, that can be describe as a set of attributes (e.g. altitude, reflectance, NDVI, etc), a binary rule can be established so that the samples set can be divided into two more homogenous data sets than the original set. This procedure will occur until the divisions lead up to each desired class of attributes. The decision rules are obtained by the definition the best discriminative function based on linear combinations of the certain attributes (Breinman et al., 1984).

With the generation of all the described images it was obtained a set of attributes that were extracted from the training and testing data sets. These sets were used to generate and choose the best decision trees, using the Gini algorithm.

2.3.3. Neural Networks

The neural networks are problem solving algorithms of the artificial intelligence that use methods and techniques inspired on historical facts and models of biological neurons and networks. These biological inspired models are extremely efficient when the pattern of classification is not a simple and trivial one (Barreto 2002). Theses networks have shown to be helpful in the resolution of problems of practical scope. Problems such as voice recognition, optical character recognition, medical diagnosis and other practical scope problems are by no means complex problems to the human brain and sensor as they are for a computer to resolve. Theses problems however can be resolved computationally through an artificial network of neurons.

Even though some researchers do not recognize the neural networks as being the general natural solution surrounding the problems of recognizing patterns on processed signals, it can be noticed that a well trained network is capable of classifying highly complex data (Kanellopolous et all 1997).

According to (Wilkinson 1997) the use of neural networks in pattern recognition and classification has grown in the last years in the field of remote sensing. A neural network needs to be capable of transforming spectral radiations into thematic maps that represent the reality.

For the interest of this work we used different types of networks. A SOM (self organizing maps) Kohonen (1990), network was used for classifying the vegetation with

supervised learning. The following described parameters used in this research on neural networks were reached through experiments and tests and limited computer power. The networks were trained and re-trained several times. Various tests were done with different network parameters aiming to reach the networks that best classified our problem of generating vegetation map. The supervised SOM had the following parameters: 31 layers with 31 neurons per layer, with variable learning rate that went from 0.5 to 1.0. The number of epochs required was of 20776 with a final quantification error of 0.1672.

A multi layer perceptron was also used for this work with the following parameters: only one hidden layer, sigmoid activation function, initial neighborhood radius of 46.25, learning rate of 0.1 and momentum of 0.5. The network was trained with 10000 iterations that lead out 95.11% of correct classifications in the training data set.

2.3.4. Soft classifiers

For (Mather 1999) the use of Fuzzy, or soft classifiers, is adequate when we want to avoid errors of classification due to ambiguity of the classes generated during the classification. When a pixel has characteristics that can include it in two or more classes, future errors of classification will occur due to this ambiguity. Fuzzy maps allow a determined pixel to be in different classes at the same time depending on the pertinence level of the pixel to each class. A fuzzy ArtMap can be generated based on the ART (Adaptavide Resonance Theory) (Carpenter et al, 1991) which is a theory that describes the biological cognitive learning of the living creatures. The ART networks were specially developed to resolve the stability-plasticity dilemma and exhibit a high degree of stability in order to preserve significant past learning, but remains adaptable enough to incorporate new information whenever it might appear (Carpenter, 1989). Fuzzy ART is a clustering algorithm that operates on vectors with fuzzy analog input patterns (real numbers between 0.0 and 1.0) and incorporates an incremental learning approach which allows it to learn continuously without forgetting previous learned states.

2.4 Training and Classification

The classification preceded as described earlier, using training samples which correspond to approximately 30 % of the total number as seen in **Table 1**. The training phase must happen to each algorithm before it can be used for classification. In the maximum likelihood algorithm training, it was considered that each pixel had the same probably of being in each class.

It was used the same training data set for the maximum likelihood, the decision tree and the also in all the kinds neural networks and fuzzy ArtMaps. For the test of theses decision trees a new set data was extracted from the original data set with values described in **Table 1,** for which the set had the same number of pixels as the training data set.

2.5 Accuracy and comparison of the generated images

As commented earlier, the main objective of this work is to verify the accuracy of these classifiers comparing them. To accomplish this, a set of accuracy samples were used as

seen in **Table 1**. With the accuracy samples a confusion matrix was generated, by which the Kappa coefficient (Colganton & Green, 1999; Tso & Mather, 2001) was extracted from. By doing this, we can compare statistically the quality of each algorithm to resolve this problem with its dataset.

3. Results

3.1. Data mining, image classification, analyses of the matrixes.

After the training phase a multivariate decision tree with the lesser possible relative cost was chosen, that is, one that has smallest possible mixture of classes on the terminal leaves (Breiman et al., 1984).

With the previous selection of the tree and its respective confusion matrix was generated using the accuracy samples. These matrixes were also generated using the maximum likelihood and for each of the type of neural classifier.

The set of temporal images obtained high Kappa coefficient values (**table 3**). This can be explained by the fact that a temporal set of images captures the phonological cycle of the vegetation.

TABLE 3 - Results of Kappa coefficient from the set of images Temporal Landsat

classification	**Max likelihood**	**Decision tree**	**MLP**	**Supervised SOM**	**Fuzzy ArtMap**
	Kappa	Kappa	Kappa	Kappa	Kappa
Values	0,9190	0,9574	0,9465	0,8043	0,9635

The classification of the Cerrado biome followed the previous steps which included classification with: MPL, Fuzzy ArtMap neural network, decision tree, and maximum likelihood from the set temporal EMT+ (**Figures 1a, 1b e 1c).** A confusion matrix was generated for the evaluation of the best Kappa coefficient **Table 4**

After applying the Landis & Koch (1977) evaluation all the classifications were all defined as excellent.

It was noticed that the Fuzzy ArtMap neural network obtained a better efficiency than all the others algorithms analyzed, thus assuring the better quality for this algorithm to classify the phytophysiognomies of the Cerrado biome.

TABLE 4 - Confusion matrix for Temporal Landsat with 96,98% of accuracy and 0,9635 of Kappa Coenfficient

Class	Water	Cropland	Bare soil	Shades	Pasture	Eucalipte plantation	Savannic	Grassland	Forests	Total
Water	1480	29	85	13	9	12	12	4	9	**1653**
Cropland	0	386	0	0	0	0	0	0	0	**386**
Bare soil	3	2	2953	0	5	11	0	1	24	**2999**
shades	2	0	0	469	0	9	11	0	0	**491**
pasture	2	1	1	0	4264	0	14	18	0	**4300**
Eucalipte plantation	3	0	23	8	0	2499	3	59	1	**2596**
Savannic	8	0	0	10	15	16	2184	72	1	**2306**
Grassland	2	0	7	0	26	83	96	7731	18	**7963**
Forests	0	2	41	0	1	0	0	15	3297	**3356**
Total	**1500**	**420**	**3110**	**500**	**4320**	**2630**	**2320**	**7900**	**3350**	**26050**

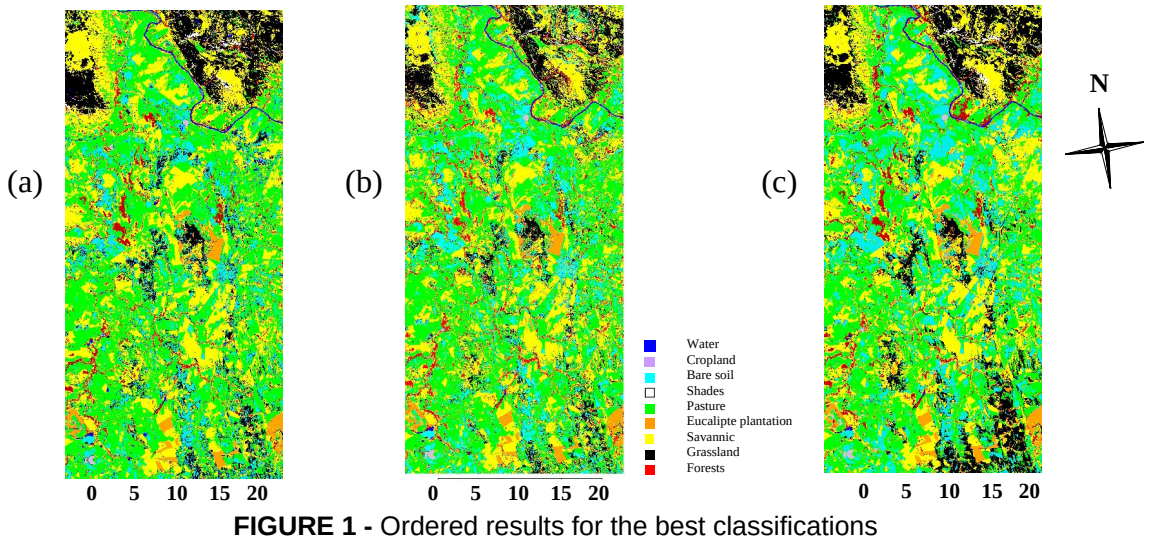

FIGURE 1 - Ordered results for the best classifications
(a) Fuzzy ArtMap neural network, (b) Decision tree (c) MLP

4. Conclusions

There are several artificial intelligence algorithms that can be used in remote sensed data to classify images and generate theme maps. All these algorithms depend in some way to the operators experience in setting up parameters of the algorithms to reach their optimal performance. When these parameters are set wisely all the algorithms work efficiently showing good overall performance, thus remembering that all these parameters should be readjusted to different data sets. The algorithms: Max likelihood, Decision tree, Decision tree, Multi layer perceptron, Fuzy ART maps showed efficiency in classifying the phytophysiognomies in the Cerrado biome. The supervised neural network using Fuzzy ARTmaps was the most efficient of the algorithms, followed by the decision tree, multy-layer perceptron and maximum likelihood.

The results show that machine learning algorithms are highly capable of mapping the Phytophysiognomies of the Brazilian Cerrado and should be highlighted that these techniques could be improved in future work so that influence of the operator should be diminished on the results.

5. References

Aguiar, A. P. D. (1991) Utilização de atributos derivados de proporções de classses dentro de um elemento de resolução de imagem (“pixel”) na classificação multiespectral de imagens de sensoriamento remoto. São José dos Campos: INPE,.

Breiman, L., Friedman, J. H., Olshen, R. A. (1984) Classication and regression tress. Belmont: Chapman & Hall,. 358 p.

Carpenter G.A, (1989) Neural Network Models for Pattern Recognition and Associative Memory. Neural Networks, 2, 243-257,.

Carpenter G. A., (1991) Crossberg, S., and Reynolds, J.H, ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network. Neural Networks, 4, 565-588,.

Carvalho, L. M. T. (2001) Mapping and monitoring forest remnants: a multi-scale analysis of spatio-temporal data. 2001. 140 p. Thesis (Doctor) - Wageningen University, Wageningen..

Colgaton, R. G., Green, K. (1999) Assenssing the accuracy of remotely sensed data: principles and practices. New York: Lewis Publishers, 137 p.

Correia, V. R. M. (1983) Estudo das medidas de qualidade para estimação de proporções de classes em elementos de resolução de imagens.. Dissertação (Mestrado) - INPE, São José dos Campos.

Crist, E. P., Cicone, R. C. (1984) A physically – based transformation of thematic mapper data – the TM Tasseled Cap. IEEE Transactions on Geocience and Remote Sensing, Los Alamitos, v. 22. n. 3, p. 256-262,.

Gaboardi, C. (2002) Utilização de imagem coerência SAR para classificação do uso da terra: Floresta Nacional do Tapajós. 137 p. Dissertação (Mestrado) - INPE, São José dos Campos.

Winkinson G. (1997) Open Questions in Neurocomputing for Earth Observation.

Instituto Nacional de Pesquisas Eespeciais. (2002) Divisão de Processamento de Imagens (INPE-DPI). SPRING, Manual do usuário [on line]. São José dos Campos.

Kouokoulas, S.; Blackbum, G. A. (2001) Introducing new indices for accuracy evaluation of classified images representing semi-natural woodland environments. Photogrammetric Engineering & Remote Sensing, Los Alamitos, v. 67, n. 4, p. 499 -510.

Landis, J. R., Koch, C. H. (1977) The measurement of observer agreement for categorical data. Biometrics, Washington, v. 33, n. 3, p. 159-174.

Lillessand, T. M., Kiefer, R. W. (1999) Remote sensing and image interpretation. 4. ed. USA: John Wiley. 724 p.

Louzada, J. N. C. (2000) Efeitos da fragmentação florestal sobre a estrutura da comunidade scarabaeidae (Insecta, coleoptera). 87 p. Tese (Doutorado) – Universidade Federal de Viçosa, Viçosa, MG.

Machado, A. B. M., Fonseca, G. A. B., Machado, R. B., Aguiar, L. M. S., Lins, L. V. (1998)Livro vermelho das espécies ameaçadas de extinção da fauna de Minas Gerais. Belo Horizonte: Fundação Biodiversitas. 608 p.

Mather, P. M. (1999) Computer processing of remotely-sensed images: an introduction. 2. ed. Nottingham, UK: Johb Wiley,. 292 p.

Mendonça, M. P., Lins, L. V. (2000) Lista vermelha das espécies ameaçadas de extinção da flora de Minas Gerais. Belo Horizonte: Fundação Biodiversitas/Fundação Zôo-Botânica de Belo Horizonte. 160 p.

Molenaar, M. (1998) An introduction to theory of Spatial object modelling for GIS. Enschede, The Netherlands: Taylor & Francis,. 246 p.

Moreira M. A., (2003) Fundamentos de Sensoriamento Remoto e Metodologias de Aplicação, 2ª Edição Revista e Ampliada Eidtora UFV 295p.

Kanellopoulos G.G., Wilkinson F.Roli, J.Austin, (1997) Neuro-computation in Remote Sensing Data Analysis,.

Kohonen, T., 1990, The Self-Organizing Map. Proceedings of the IEEE, 78: 1464-80.

Oliveira, L.T.; (2004) Fusão de imagens de sensoriamento remoto e mineraçao de dados geográficos para mapear as fitofisionomias do Bioma Cerrado.. 131p. (CDD – 621.3678 – 526.982) Dissertação (Mestrado em Manejo Ambiental)– UFLA. Lavras. 2004.

Ribeiro, J. F., Walter, B. M. T. (1998) Fitofisionomias do bioma Cerrado. In: Sano, S., Almeida, S. P. (Ed.) Cerrado: ambiente e flora. Planaltina: EMBRAPA-CPAC,. p. 89-169.

Schowengerdt, R. A. (1997) Models and metods for image processing. 2. ed. New York: Academy Press,. 522 p.

Shimabukuro, Y. E. (1987) Shade images derived from linear mixing models of multispectral measurements of forested areas.. Dissertation (Doctor of Phylosophy) - Colorado State University, Fort Collins.

Tso, B., Mather, P. M. (2001) Classification Methods for remotely sensed data. New York: Taylor & Francis, 332 p.

Oliveira-Filho, A. T. , Ratter J. A., (2002) The Cerrados of Brazil: ecology and natural history of a neotropical savannah, Columbia University Press Publishers, New York Chischester, West Sussex p. 91-121

Jensen J. R. 2000, Remote Sensing of the enviroment: An Earth resource perspective, Pretince Hall Series in geographic information science, Upper SaddleRiver, New Jersey 07458